DCI/RM-80-1951
15 September 1980

MEMORANDUM FOR: Director, HUMINT Tasking Office

25X1 FROM: [redacted]
Program Assessment Office

VIA: Director, Program Assessment Office

SUBJECT: DoD's Intelligence Report Evaluation Program--A Statistical Review

REFERENCES:
A. NFIP and Resource Guidance FY 82-86 (9 May 80, DCI/RM/3275-80)

B. Army Clandestine HUMINT--A Review (5 Mar 80, DCI/RM 80-2001, Attachment II)

C. DIA Response to CTS/HTO Questions (31 July 80)

The DoD Intelligence Report (IR) evaluation program was developed to reflect the degree to which DoD Human Source reporting meets the requirements levied upon it. The program calls for roughly 20% of all IRs to be evaluated. Some IRs are automatically evaluated due to the collection requirements that drive them; others are evaluated at the collector's initiative, while still others are evaluated at the initiative of DoD analysts.

DoD analysts provide an IR evaluation by subjectively categorizing the value of an IR as with "high", "moderate", "low", "none" or "cannot judge".

25X1 This statistical review evaluates the soundness of DoD's IR evaluation program. [redacted]

Background:

Statistically, samples are selected from a larger population according to some rule or plan. Generally, samples are obtained by one of two methods; those selected by some form of subjective judgment, and those selected according to some chance mechanism (such as random sampling).

25X1

25X1

[redacted]

[redacted]

A good sample is one from which generalizations to the population can be accurately and precisely made. To generalize from a sample to a population, the laws of mathematical probability must apply--random sampling assures that these laws do apply. For this reason random samples are preferred over judgment samples.

To generalize accurately and precisely from a sample to a population, the uncertainties in the sample must be understood. There are two components of sample uncertainty: reliability (numerical precision) and validity (accuracy or realism). Reliability is controlled for the most part by sample size, and can be calculated from the data at hand. Validity, however, cannot be judged from the data and can be controlled only before sampling through sound experimental design. [redacted] 25X1

Discussion:

DoD's sample size of roughly 20% provides for sufficiently precise estimates, IF THE SAMPLE IS VALIDLY CHOSEN. The percentage of IRs rated as having high value, for example, are precise to better than ±3% (95% Confidence Internal) based on the 20% sampling (See Appendix). In fact, a sample as small as 500 evaluations, if chosen properly, will provide precision to better than ±5% (95% Confidence Internal).

It must be noted parenthetically that the precision of sample estimates is proportional to the number of IRs sampled and not to the percentage of IRs sampled. Reference C states that samples were taken from each of some 120 individual collection entities. Care must be taken when examining separately each of these collection entities since their sample sizes may be quite small. On the average, one would expect the precision of estimates within a collection entity to be on the order of ±10-20% (95% Confidence Interval).

However, it is not insufficent reliability but insufficient validity that undermines DoD's evaluation program. There are three primary causes of invalidity:

(1) Systematic errors. According to Reference C, there is a tendency to initiate evaluations of high-or-low-value reports at the expense of reports rated moderate in value. This practice results in the systematic elimination of a portion of the population and a consequent bias to inferences made from the sample. Reference B surfaces another source of systematic error: the inordinate number of high evaluations that upon closer examination appear to have been unwarranted. The effects of such systematic overrating cannot be removed through statistical analysis and thus further undermine the validity of the inferences drawn from the sample.

25X1

(2) Mismatch between sample and population. Reference B also isolates a serious mismatch between the sample and the population it purports to represent--the sample was taken primarily from the population of mid-level DoD analysts while inferences are drawn about the population of consumers (policymakers and senior analysts both inside and outside DoD). Since the value of a report to mid level analyst appears to be different from the value of the same report to other consumers (Reference B), one must seriously question the use to which DoD's summaries can be put.

Furthermore, DoD's evaluation sample does not appear to match the total IR population in several other respects. The sample was not randomly chosen (i.e., each report did not have an equal chance of being evaluated), thus invalidating the mathematical basis for making inferences. As noted before, judgment sampling is not random, and according to Reference C, "analyst initiative" evaluations are often intentionally biased to "reduce the ... IRs which ... are evaluated as being of low or no value." Likewise, it is not clear that special and initiative evaluations are representative of the total IR population, since they represent reports of some special, not random, interest.

Failure to attend to the representativeness of the sample can lead to serious underestimates of uncertainty and consequent overoptimism about the stability and realism of population inferences. And estimates for which the accuracy is unknown can be quite misleading.

(3) Correlated evaluations. If one analyst evaluates a disappropriate share of reports and has a tendency to rate reports higher or lower than other analysts, his evaluation may speciously inflate (or deflate) the estimated worth of IRs. His evaluations are said to be correlated, and correlated evaluations lower the validity of an analysis. Likewise, if several evaluations are performed on a single requirement (or similar requirements), there is again the tendency for such correlated evaluations to artificially alter population estimates. There is potential for such correlated evaluations in "analyst initiative" reporting. [redacted]

25X1

25X1

25X1

Conclusions:

- If the intent is to understand the value to consumers of IRs as a whole, mandatory evaluation must be randomly assigned to 10% or so (depending upon the accuracy desired) of all reporting to match sample to population and to provide for sufficient reliability. Furthermore, since mid-level analysts provided evaluations from their own perspective, results will be valid only for these analysts. Inferences about other consumers are invalid unless it can be shown that the attitudes and perspective of mid-level analysts are like those of the other consumers.

- "Initiative" and specially requested evaluations, while they may be useful for other purposes, should not be included in the data analysis due to their systematic biases and potential for correlated evaluations.

- The assertion in Reference B that the Intelligence Community "cannot rely upon such evaluations for an objective view of the worth of the reporting" appears to be based on an invalidating mismatch between sample and population.

- The violation of such fundamental laws of validity renders the DoD Evaluation of questionable value for estimating the worth of intelligence reporting to consumers. [redacted]

25X1

[redacted]

25X1

25X1

APPENDIX. Statitical Foundation for Estimates of Precision.

DoD defines the value of an IR as either "high", "moderate", "low", "none" or "cannot judge". These categories form a well-defined statistical population known as a multinomial population. When samples are randomly placed into multinomial categories, the percentage of the total sample falling in each category can easily be calculated. The variance (a measure of precision) of each percentage, P, is defined as: Variance = [P(100-P)] ÷ N, where N is the total sample size. For example, if 70% of 2,000 evaluations are rated as "moderate" in value, the precision of this 70% is given by: [70(30)] ÷ 2000 = 1.05. A 95% Confidence Internal is approximated by twice the square root of this number, or about 2. Therefore, the 70% is precise to within ±2% (at a 95% level of confidence). In other words, if this evaluation were repeated 100 times, one would expect the proportion of "moderate" ratings to be between 68% and 72% 95 times, and outside that range only 5 times. [redacted] 25X1

SUBJECT: DoD's Intelligence Report Evaluation Program--A Statistical Review

Distribution: (DCI/RM-80-1951)

Copy 1 - D/HTO
2 - D/PAO
25X1 3 - PAO [redacted]
4 - PAO
5 - HTO
6 - HTO
7 - PBO
8 - PAO Subject
9 - PAO Chrono
10 - RM Registry
11 - CT Registry

25X1 DCI/RM/PAO: [redacted] 8 Sep 80)

Next 1 Page(s) In Document Exempt

Approved For Release 2005/03/24 : CIA-RDP83M00171R000700030001-7